OUTLINE OF THE PROPOSED TOPIC OF RESEARCH

Candidate

Name of the Candidate	: Pankaj Jadwal
ID No.	: 2013PhDENGG005
Place of Research Work and Organization	: JK Lakshmipat University, Jaipur (JKLU)

Proposed Supervisor

Name	: Dr. Sonal Jain
Qualification	: Ph.D.
Designation	: Associate Professor,
	: Institute of Engineering and Technology
Organization	: JKLU, Jaipur

Proposed Topic of Research

Study of clustering technique for decision making in banking sector

1. Introduction

Banking sector

Introduction

India is one of the top 10 economies globally, with vast potential for the banking sector to grow. The last decade witnessed a tremendous upsurge in transactions through ATMs, and Internet and mobile banking. In 2014, the country's Rs 81 trillion (US$ 1.34 trillion) banking industry is set for a greater change. Two new banks have already received licences from the government. Furthermore, the Reserve Bank of India's (RBI) new norms will provide

incentives to banks to spot potential bad loans and take corrective steps that will curb the practices of rogue borrowers.

The Indian government's role in expanding the banking industry has been significant. Through the Financial Inclusion Plan (FY 10–13), banking connectivity in the country increased more than three-fold to 211,234 villages in 2013 from 67,694 at the beginning of the plan.

Banks are also looking at new ways to attract customers. In September, 2013, ICICI bank leveraged the popularity of the social platform, and launched its Facebook banking service, Pockets. The service enables customers to transfer funds and pay bills from within the website. (Damji, B. H. ,2009)

Market Size

The revenue of Indian banks increased four-fold from US$ 11.8 billion to US$ 46.9 billion during the period 2001–2010. In the same period, the profit after tax increased from US$ 1.4 billion to US$ 12 billion.

In 2012–13, Indian banks had 170 overseas branches (163 in 2011–12) while foreign banks had 316 branches in India (309 in 2011–12).

Credit to housing sector grew at a compound annual growth rate (CAGR) of 11.1 per cent during the period FY 2008–13. Total banking sector credit is expected to grow at a CAGR of 18.1 per cent (in terms of INR) to touch US$ 2.4 trillion by 2017. (Demetriades, P. O., & Luintel, K. B. ,1996)

Challenges in banking sector

N Ravichandran (Professor, IIMA): The economic reforms initiated by the Government of India roughly about a decade ago have changed the landscape of several sectors of the Indian economy. The Indian banking sector is no exception. This sector is going through major changes as a consequence of economic reforms. The changes affect the ownership pattern of banks, availability of funds, the cost of funds as well as opportunities to earn, range of services (fee-based andfund-based), and management of priority sector lending. As a

consequence of liberalization in interest rates, banks are operating on reduced spread. Development financial institutions would have a lesser impact on the Indian economy. Consumerism is here to stay. Non-banking products like insurance would be a tremendous opportunity.The economic reforms have also generated new and powerful customers (huge Indian middle class) and new mix of players (public sector units, private banks, and foreign banks). The emerging competition has generated new expectations from the existing and the new customers.There is an urgent need to introduce new products. Existing products need to be delivered in an innovative and cost-effective way by taking full advantage of emerging technologies.The new rules of competition require recognition of the importance of consumers and the necessity to address the needs through innovative products supported by new technology.As a consequence of competition, the managerial challenges include market segmentation, product positioning,innovative delivery channels, cross-selling, etc. At an organization level, elaborate systems need to be evolved to manage, assess, and contain risk (including portfolio, client, and exchange rate). The banks may have to reorient their resources in the form of reorganized branch networks, reduced manpower, dramatic reduction in establishment cost, honing the skills of the staff, and innovative ways of attracting talented managerial pool.The Government of India and the Reserve Bank of India (RBI) on their part would strengthen the existing norms in terms of governing and directing the functioning of these banks. Banks needs to strengthen their audit function. They would be evaluated based on their performance in the market place. It is in this context that we have invited the chief executive officers of Indian banks to respond to the issues mentioned earlier. (Kamath, K. V., Kohli, S. S., Shenoy, P. S., Kumar, R., Nayak, R. M., Kuppuswamy, P. T., & Ravichandran, N. 2003)

In his (Jalan, B. ,2002) inaugural address last year, He had indicated a vision for Indian banking in the new millennium – that of a vibrant, internationally active banking system, drawing upon its innate strengths and comparative advantages to make India a major banking centre of the world. He had pointed out then that, while it may take up to 10 or even 15 years to achieve this vision, the time to begin was now. Recent developments have only served to bring forward the urgency attached to embarking upon this quest. Even as we do so, it is necessary to recognise that, in view of recent global developments and the economic slowdown, the progress towards this goal would call for even greater effort and determination. In this context, the theme chosen for this year's Conference *i.e.*, "Indian Banking: Paradigm Shift" is most timely as it provides an opportunity to deliberate on the

new challenges ahead, and the action that we must take to manage them. He is happy to be a part of these deliberations and to deliver the inaugural address to the 23 rd Conference of Bank Economists .(Jalan, B. ,2002).

Clustering

Introduction

The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups. Distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. (Jiawei, H., & Kamber, M. ,2001)

In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery. Clustering is also called data segmentation in some applications because clustering partitions

large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers (values that are "far away" from any cluster)may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, exceptional cases in credit card transactions, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity. As a data mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as characterization, attribute subset selection, and classification, which would then operate on the detected clusters and the selected attributes or features. (Jiawei, H., & Kamber, M. ,2001)

Data clustering is under vigorous development.Contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in data mining research.

As a branch of statistics, cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS. In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. Active themes of research focus on the scalability of clustering methods, the effectiveness of methods for clustering complex shapes and types of data, high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases. (Jiawei, H., & Kamber, M. ,2001) Clustering is a challenging field of research in which its potential applications pose their own special requirements.

Requirements of clustering in data mining:

Scalability:

Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of objects. Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed. Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data.

However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types. Discovery of clusters with arbitrary shape: Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape. Minimal requirements for domain knowledge to determine input parameters: Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often difficult to determine, especially for data sets containing high-dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.

Incremental clustering and insensitivity to the order of input records:

Some clustering algorithms cannot incorporate newly inserted data (i.e., database updates) into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data. That is, given a set of data objects, such an algorithm may return dramatically different clustering depending on the order of presentation of the input objects. It is important to develop incremental clustering algorithms and algorithms that are insensitive to the order of input.

High dimensionality:

A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in high dimensional space is challenging, especially considering that such data can be sparse and highly skewed. (Jiawei, H., & Kamber, M. ,2001)

Constraint-based clustering:

Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic banking machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster. A challenging task is to find groups of data with good clustering behavior that satisfy specified constraints. (Jiawei, H., & Kamber, M. ,2001)

Interpretability and usability:

Users expect clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied to specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and methods. (Jiawei, H., & Kamber, M. ,2001)

With these requirements in mind, our study of cluster analysis proceeds as follows. First, we study different types of data and how they can influence clustering methods. Second, we present a general categorization of clustering methods. We then study each clustering method in detail, including partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. We also examine clustering in high-dimensional space, constraint-based clustering, and outlier analysis.

Clustering is an unsupervised operation. It is used where you wish to find groupings of similar records in your data without any preconditions as to what that similarity may involve. Clustering is used to identify interesting groups in a customer base that may not have been recognized before. For example, it can be used to identify similarities in customers' telephone usage, in order to devise and market new call services.

Clustering is usually achieved using statistical methods, such as a k-means algorithm, or a special form of neural network called a Kohonen feature map network. Whichever method is used, the basic operation is the same. Each record is compared with a set of existing clusters, which are defined by their 'centre'. A record is assigned to the cluster it is nearest to, and this in turn changes the value that defines that cluster. Multiple passes are made through a data set to re-assign records and adjust the cluster centers until an optimum solution is found.

Looking for clusters amongst supermarket shoppers, for example, may require the analysis of many factors, including the number of visits made per month, the total spend per visit, spend per product category, time of visit and payment method. (Jiawei, H., & Kamber, M. ,2001)

Types of Clustering Methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, The known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general , clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar, however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods , the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisivemethods begin with all objects in a single cluster and at each of N-1 steps divides some clusters into two smaller clusters, until each object resides in its own cluster.(Ali, R., Ghani, U., & Saeed, A. ,1998)

Some of the important Data Clustering Methods are described below

1 **Partitioning Methods**

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that

occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient. (Ali, R., Ghani, U., & Saeed, A. ,1998)
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order O(NlogN) for order O(logN) clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run

2 Hierarchical Agglomerative methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects. (Ali, R., Ghani, U., & Saeed, A. ,1998)
3. If more than one cluster remains , return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below

- In the second matrix approach , an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.
- The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement. (Ali, R., Ghani, U., & Saeed, A. ,1998)

3 The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

4 The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

5 The Group Average Method

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster.

6 Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document. (Ali, R., Ghani, U., & Saeed, A. ,1998)

Applications of Clustering

Data clustering has immense number of applications in every field of life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So the history of data clustering is old as the history of mankind.

In computer field also, use of data clustering has its own value. Specially in the field of information retrieval data clustering plays an important role. Some of the applications are listed below.

1 Similarity searching in Medical Image Database

This is a major application of the clustering technique. In order to detect many diseases like Tumor etc, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognized.

We have clusters of images of different parts of the body. For example, the images of the CT Scan of brain are kept in one cluster. To further arrange things, the images in which the right side of the brain is damaged are kept in one cluster. The hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. In this form a large database of images is maintained using the hierarchical clustering.

Now when a new query image comes, it is firstly recognized that what particular cluster this image belongs, and then by similarity matching with a healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then the image is sent to that specific cluster and matched with all the images in that particular cluster. Now the image with which the query image has the most similarities, is retrieved and the record associated to

that image is also associated to the query image. This means that now the disease of the query image has been detected. (Ali, R., Ghani, U., & Saeed, A. ,1998)

Using this technique and some really precise methods for the pattern matching, diseases like really fine tumor can also be detected.

So by using clustering an enormous amount of time in finding the exact match from the database is reduced.

2 Data Mining

Another important application of clustering is in the field of data mining. Data mining is defined as follows.

Definition1: "Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques." [5]

Definition2: Data mining is a "knowledge discovery process of extracting previously unknown, actionable information from very large databases." [5]

Use of Clustering in Data Mining: Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

For example, a company that sales a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving less sale then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process. (Ali, R., Ghani, U., & Saeed, A. ,1998)

3 Windows NT

Another major application of clustering is in the new version of windows NT. Windows NT uses clustering, it determine the nodes that are using same kind of resources and accumulate them into one cluster. Now this new cluster can be controlled as one node.

The detailed study of clustering in Windows NT is given below as the Case Study.

2. Review of literature

This paper shows the first results of a statistical methodology to cluster the business models of a large samples of European banks and to identify peer group of selected banks. This analysis is based on publically available data from banks financial statements. This paper defines 5 variables to define the business model of the bank. A k centroid Clustering method is used for assigning different bank to different clusters. (Ferstl, R., & Seres, D. ,2012)

This study(Ashton, J. K., & Hudson, R. S. 2008) forwards an explanation and empirical investigation of price clustering in retail banking markets. It is proposed that price or interest rate clustering forms in retail markets as firms wish to maximise returns from customers, some of whom have difficulties in recalling and processing price information. This theory is developed and tested using a dataset of retail interest rates from the UK which enables interest rate clustering to be viewed in both lending and investment markets, and at different levels of financial involvement. It is found that interest rate clustering occurs in a manner consistent with firms maximising returns from customers. These findings are viewed to be a key policy concern for financial regulators and firms concerned with consumer protection.

Hard clustering and fuzzy clustering analysis is the basis of the construction of many classifications and systems with its main focus on planning and dividing data into many subsets according to certain rules.Due to the practical function of clustering analysis, many researchers thus proposed different clustering algorithms to be used by researchers around the world. In this article, (Pan, W. T. ,2010) Fuzzy sammon Mapping method is implemented to perform clustering effect and classification capability analysis on these frequently used clustering algorithms. From the result of test data of an investigation of banking service satisfaction, GK Cluster algorithm was found to have very good clustering effect; however,

as for classification capability, hard clustering analysis method has proved to be the better approach amongst the two.

Recent literature has focused on the study of systemic risk in complex networks. It is clear now, after the crisis of 2008, that the aggregate behaviour of the interaction among agents is not straightforward and it is very difficult to predict. Contributing to this debate, this paper (Tabak, B. M., Takami, M., Rocha, J., Cajueiro, D. O., & Souza, S. R. ,2014) shows that the directed clustering coefficient may be used as a measure of systemic risk in complex networks. Furthermore, using data from the Brazilian interbank network, we show that the directed clustering coefficient is negatively correlated with domestic interest rates.

This paper (Tao, L., Liu, X., & Chen, Y. ,2013) presents a hybrid model to conducting performance measurements for Internet banking by using data envelopment analysis (DEA) and axiomatic fuzzy set (AFS) clustering. For each bank, DEA is applied to select a best combination by computing an aggregated efficiency score based on outputs, such as web metrics and revenue; and inputs, such as equipment, operation cost and employees. Then AFS clustering method is used to classify the best combination into different groups and provide the definitely semantic interpretations for the evaluation results. Identification of operational fitness and business orientation of each firm, in this way, will yield insights into understanding the weaknesses and strengths of banks, which are considering moving into Internet banking.

In present scenario, high performance cluster-based web server is needed to be deployed by banking services to fit the ever – increasing demands of the online banking users. In the last few years, the increased pressure of high-online banking users have overloaded the existing web clusters and thus they fail to supply better services to all the online users resulting in unexpected long delays. To handle this situation, Object Oriented approach needs to be implementing in the web clusters. This research (Dev, H., & Mishra, S. K. ,2013) proposes how to improve the performance of the data clustering by using the object - oriented approach.

To solve the problem of credit rating of loan application, this paper (Xiang, J. P., Qiao, S. J., & Hu, J. ,2011) introduces the basic concepts and preliminaries of clustering analysis employed in this study to handle the problem of estimation on the loan grade of clients and presents two kinds of clustering algorithms for solving this problem. One clustering approach is based on credit data, called δ-kmeans,and the other approach can be applied to high dimensional credit data,called ASC.Extensive experiments were conducted to compare the performance between

the proposed two algorithms.The experiments show that δ-kmeans algorithm achieves better results in credit risk control,but the ASC algorithm is more effective than traditional k-means algorithms and the Coweb algorithm in clustering high dimensional credit data.In addition,this paper analyzes the dynamics of cluster analysis on bank credit data in terms of the k-means algorithm.

Credit risk is a key issue in the bank loans,therefore,the prediction of credit risk of offering loans to farmers is very important in the business of micro agricultural loans.The author adopts the fuzzy clustering to conduct a soft classification of farmers,searches for a reasonable classification by using F test and constructs a model for predicting the credit risk of micro agricultural loans.The experiment results indicate that the method achieves good results and has bright application prospects.(Yan-ling, W. E. I. ,2011)

Cluster analysis, as an exploratory technique, by gathering together those credit institutions sharing similar features in terms of financial intermediation activity, proves to be a complementary tool for the peer group analysis, accomplished at the off-site supervision level. The aim of study was to include a representative sample of Romanian credit institutions into smaller, homogenous clusters, in order to assess which credit institutions have similar patterns according to their risk profile and profitability. They found that, over the period 2004-2006, the clusters remained relatively stable in terms of similarity of exposure to risks and profitability. (Dardac, N., & Boitan, I., 2009)

In this paper (Vagizova, V., Lurie, K., & Ivasiv, I. , 2014) clustering method is used to identify the business models of the credit institutions in the formed groups, to consider the interconnection of the banks in every cluster with the real sector of the economy and suggested possible development ways of this interaction. The hypothesis of the institutional misalignment in the Russian banking system has been confirmed, and they have traced the interaction problem sources between the banking sector and the real sector of the economy during the analysis of the bank business models existing at present. They plan the further development of effective monetary resources management tools for the insurance of the interaction between the banking sector and the real sector of the economy.

Market segmentation is one of the most important area of knowledge-based marketing. In banks, it is really a challenging task as data bases are large and multidimensional. In the paper they consider cluster analysis, which is the methodology, the most often applied in this area. They (Zakrzewska, D., & Murlewski, J. ,2005, September) compare clustering algorithms in cases of high dimensionality with noise. They discuss using three algorithms: density based DBSCAN, k-means and based on it two-phase clustering process. They compare algorithms concerning their effectiveness and scalability.

The unprecedented growth of competition in the banking technology has raised the importance of retaining current customers and acquires new customers so that is important analysing Customer behaviour, which is base on bank databases. Analysing bank databases for analysing customer behaviour is difficult since bank databases are multi-dimensional, comprised of monthly account records and daily transaction records. Few works have focused on analysing of bank databases from the viewpoint of customer behavioral analyze. This study ((FARAJIAN MOHAMMAD, A. L. I., & MOHAMMADI, S. ,2010) presents a new two-stage frame-work of customer behaviours analysis that integrated a K-means algorithm and Apriori association rule inducer. The K-means algorithm was used to identify groups of customers based on recency, frequency, monetary behavioural scoring predicators; it also divides customers into three major profitable groups of customers. Apriori association rule inducer was used to characterize the groups of customers by creating customer profiles. Identifying customers by a customer behaviour analysis model is helpful characteristics of customer and facilitates marketing strategy development.

In today's businesses, achieving customers satisfaction have critical role in organization's goals. On the other hand, all of customers hasn't equal share in profitability of organization. Therefore, identification key customers will be more sensitive. Calculate the lifetime value assist organizations to rank customers based on their contribution to profitability. The purpose of this paper is introduced a model to calculate customer lifetime value (CLV) based on LRFM customer relationship model which consists of four dimensions: relation length (L), recent transaction time (R), buying frequency (F), and monetary (M) in banking services. We proceed with this clustering analysis to classify customers in order to set marketing strategies. In this research, they (Alvandi, M., Fazli, S., & Abdoli, F. S. ,2012) used crisp method and implemented them on real data from an Iranian state bank. Validity of clustering process analyzed with R-Squared index. The results show nine cluster patterns between customers.

Finally, in terms of this clustering, they proposed customer strategies. Thus, this study considers useful for customer relationship management.

Islamic bank in Malaysia was established since 1983 and the development of Islamic bank is increasingly challenging with more conventional banks involved and offered banking products and services based on Islamic principles. Therefore, it is crucial for Islamic bank to better position themselves in order to undertake and respond enthusiastically to the robustness of the dual banking environment. This paper (Januri, S. S., Ismail, N., & Idris, A. R. .2011, June) analysed the factor considered important by academicians in public higher institution in selecting Islamic banking. The study revealed that religious value appears to be the most important factor. Other factors perceived to be important include trustworthiness, product variety, service facility and service efficiency. A cluster analysis indicates that customers can be divided into three clusters with distinctive demographic and factors for each cluster. Overall, members for all clusters tend to choose an Islamic banking because of religious value and trustworthiness factors. The results will benefit banking management because they provide information to the importance of the selection criteria.

Due to this globalization, the banks struggle to gain an edge over others. To overcome this, banks has rapidly increased the ability to generate, manipulate and storage of data of the customers because the information contained in this data can be highly useful. The E-banking industry around the globe has undergone a remarkable change in the way business is conducted. The banking industry has realized the need of the techniques like data mining, web mining, etc which can help them to compete in the market and to make the prediction and decision for the industry for growth. Banks can use data mining tools like clustering on customers profile. This paper (Bansal, K., & Bohra, A. 2013, April) provides an overview of the concept of K- mean clustering algorithm and highlights the applications of e-banking to increase the customers satisfaction.

A cluster is a group of objects that have almost the same characteristics. Clustering evaluation is a big problem in this field of studies. Several attempts were made in order to overcome the problem. Three ways exist to try to approach the problem: internal, external and relatively. The internal approach use measures that are only related with the data in study. The external way use labels to try to solve the problem. The relatively way use some kind of criterion to choose the best division. This last case is the one author is going to work

with, using the marketing idea about what a good cluster is. Marketing has as final goal attract and retain customers. For that it is important to know and satisfy them. Segmentation helps in this goal. Six criterions define what a good cluster is in the marketing perspective: identifiability, accessibility, substantiality, responsiveness, stability and actionability. However these criterions are only formal defined. The final goal of this (Loureiro. A. ,2011)work is to try to find a way to transform these formal criterions in measures that allow us to evaluate the problem in a quantitative way.

3. Rationale

Clustering is tool which have been applied in so many areas like educational, medical ,spatial ,economical and banking sector. Literature survey reveals that applications of clustering have been used in many sectors with enough quantum but still gap exists when banking sector is considered. Paving a way to continue research field ahead. Banking is the field in which clustering have been applied for improving customer bank relationship and creditability of the bank. Clustering may help them to compete in the market and to make the prediction and decision for the banking sector for growth. It may clusterize the customers on the basis of different parameters (like type of customer, need of customer etc.).Clustering concept might be applied on banks so that appropriate decisions regarding funds or facilities may be taken by the RBI.Clustering may be applied on granting loan so that applicants may be classified in the clusters and it may be decided that loan should be given or not. Hierarchical clustering would help in generating clusters from clusters.

4. Objective

1. Study of banking sector where decisions are to be made

2. Clustering may be applied to Banking sector for getting optimal results from a group of feasible results in the decision making process.

3. Applications of clustering for decision making filtering techniques(For eg. Housing loan, Bank customer interaction etc.)

4. Clustering for promoting E-Banking Services

5.Clustering may be applied for grouping of different ATMs on the basis of different parameters so that decision can be taken regarding number of machine in each cluster.

5. Hypotheses and Data Set Information

a. Problem statement/ hypotheses

This research aims to apply clustering techniques for helping Banking sector in decision making process in several domains

b. The Sample

Data set of banking sector is taken and data is taken from UCI repository(S. Moro, R. Laureano & P. Cortez;2010)

There are two datasets:

1) bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).

2) bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.

Data Set 2 is taken as sample data.

c. Data Collection Tools:

1. Data set of banking sector is taken and data is taken from UCI repository(S. Moro, R. Laureano & P. Cortez , 2010).

6. Possible Outcomes

In view of the proposed plan of the research, it is anticipated that the following possible outcomes will emerge out of the research:-

1. Bank customer relations will be improved by improving the decision support system of Banking system with the help of clustering.

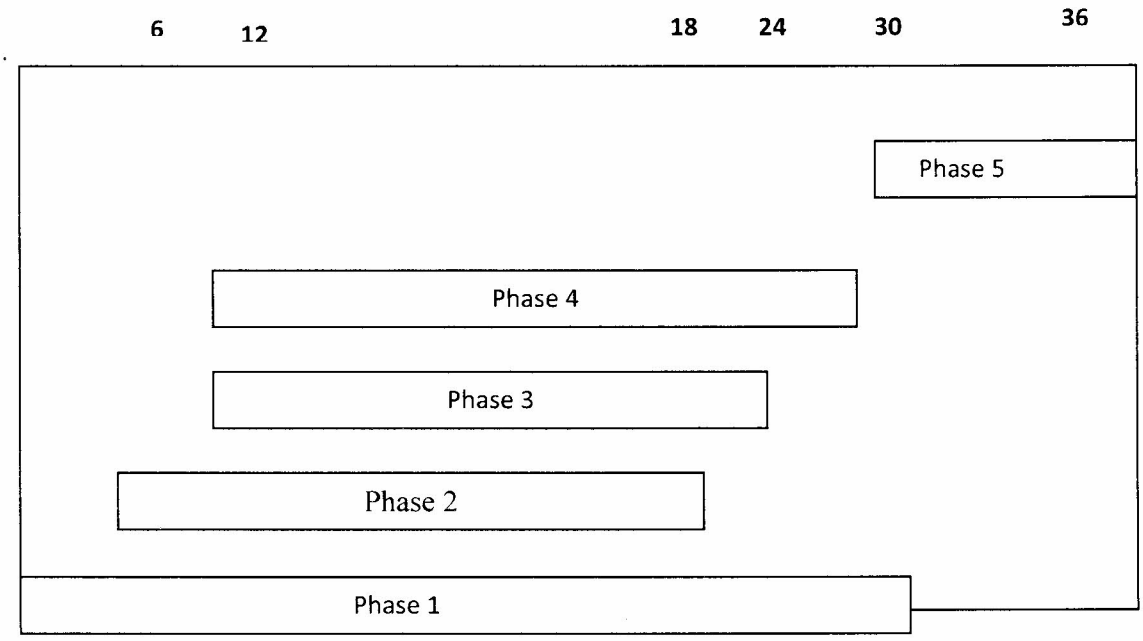
2. Different aspects of bank(like loans,creditability etc.) can be improved via clustering.

3.Bank will pay more concentrate on the cluster which is useful for them.So it will save their time and money.

4.All domains and aspects where clustering may be applied for improving the decision support system will be applied.

7. Work Plan

The work-plan for the above-mentioned seven activities is shown in figure 1



Phase 1:Literature Review

Phase 2: Data collection and preprocessing of data

Phase 3: Experimental set up and applying clustering tool on preprocessed data.

Phase 4: Analysis of results

Phase 5: Conclusion and Thesis Writing

References

Ali, R., Ghani, U., & Saeed, A. (1998). Data clustering and its applications.*Available at the web address: http://members. tripod. com/asim_saeed/paper. htm.*

Alvandi, M., Fazli, S., & Abdoli, F. S. (2012). K-Mean clustering method for analysis customer lifetime value with LRFM relationship model in banking services. *International Research Journal of Applied and Basic Sciences*, *3*(11), 2294-2302.

Ashton, J. K., & Hudson, R. S. (2008). Interest rate clustering in UK financial services markets. *Journal of Banking & Finance*, *32*(7), 1393-1403.

Bansal, K., & Bohra, A. (2013, April). K-Mean Clustering Algorithm Implemented To E-Banking. In *International Journal of Engineering Research and Technology* (Vol. 2, No. 4 (April-2013)). ESRSA Publications.

Boschma, R. A., & Wenting, R. (2010). Spinoffs and M&A as drivers of spatial clustering: the evolution of the Dutch banking sector in the Amsterdam region in the period 1850–1993.

Damji, B. H. (2009). Banking Sector In India. *Economic Survey*, *42*, 15-000.

Dardac, N., & Boitan, I. (2009). A Cluster Analysis Approach for Bank's Risk Profile: The Romanian Evidence. *European Research Studies*, *12*(1).

Demetriades, P. O., & Luintel, K. B. (1996). Financial development, economic growth and banking sector controls: evidence from India. *The Economic Journal*, 359-374.

Dev, H., & Mishra, S. K. (2013). Improving the Performance of Banking Sector by Using Clustering Method: An Object–Oriented Approach.

FARAJIAN MOHAMMAD, A. L. I., & MOHAMMADI, S. (2010). MINING THE BANKING CUSTOMER BEHAVIOR USING CLUSTERING AND ASSOCIATION RULES METHODS. *INTERNATIONAL JOURNAL OF INDUSTRIAL ENGINEERING AND PRODUCTION RESEARCH (IJIE) (ENGLISH).*

Ferstl, R., & Seres, D. (2012). Clustering Austrian banks' business models and peer groups in the European Banking Sector. *Oesterreichische Nationalbank, Financial Stability Report.*

Jalan, B. (2002). Indian banking and finance: managing new challenges. *Speech Delivered at the*.

Januri, S. S., Ismail, N., & Idris, A. R. (2011, June). Clustering the customers on the selection of Islamic banking in East Coast Malaysia. In *Data Mining and Optimization (DMO), 2011 3rd Conference on* (pp. 259-265). IEEE.

Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, *5*.

Kamath, K. V., Kohli, S. S., Shenoy, P. S., Kumar, R., Nayak, R. M., Kuppuswamy, P. T., & Ravichandran, N. (2003). Indian banking sector: Challenges and opportunities. *Vikalpa*, *28*(3), 83-100.

Loureiro, A. (2011). Quantitative evaluation of clustering for segmentation: an application to the banking sector.

Mäenpää, K. (2006). Clustering the consumers on the basis of their perceptions of the Internet banking services. *Internet Research*, *16*(3), 304-322.

Pan, W. T. (2010). Combining fuzzy sammon mapping and fuzzy clustering approach to perform clustering effect analysis: Take the banking service satisfaction as an example. *Expert Systems with Applications*, *37*(6), 4139-4145.

Sarkar, J. (1999). India's Banking Sector: Current Status, Emerging Challenges, and Policy Imperatives in a Globalized Environment. *James A. Hanson, and Sanjay Kathuria, eds.*

Tabak, B. M., Takami, M., Rocha, J., Cajueiro, D. O., & Souza, S. R. (2014). Directed clustering coefficient as a measure of systemic risk in complex banking networks. *Physica A: Statistical Mechanics and its Applications*, *394*, 211-216.

Tao, L., Liu, X., & Chen, Y. (2013). Online banking performance evaluation using data envelopment analysis and axiomatic fuzzy set clustering. *Quality & Quantity*, *47*(2), 1259-1273.

Vagizova, V., Lurie, K., & Ivasiv, I. (2014). Clustering of Russian banks: business models of interaction of the banking sector and the real economy.

Xiang, J. P., Qiao, S. J., & Hu, J. (2011). The method based on clustering analysis to evaluate the credit rating of loan application. *Journal of Yunnan University Natural Sciences Edition*, *33*(6), 639-643.

Xionga, X., & Zhang, W. Optimal Clustering of Off-site Commercial Banking Regulation Based on Self-Organizing Feature Map1.

Yan-ling, W. E. I. (2011). Prediction of Credit Risk of Micro-loans to Farmer by Using Fuzzy Clustering Based on F Test [J]. *Journal of Anhui Agricultural Sciences*, *1*, 217.

Zakrzewska, D., & Murlewski, J. (2005, September). Clustering algorithms for bank customer segmentation. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on* (pp. 197-202). IEEE.

Zhang, J., Zhao, P., Shang, L., & Wang, L. (2009, August). Web usage mining based on fuzzy clustering in identifying target group. In *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on* (Vol. 4, pp. 209-212). IEEE.